risti
Revista Ibérica de Sistemas e Tecnologias de Informação
Iberian Journal of Information Systems and Technologies

Recebido/Submission: 05/10/2021
Aceitação/Acceptance: 03/01/2022

Mitigação dos Riscos à Privacidade através da Anonimização de Dados

Juliano Rodrigues Ferreira[1], João Alberto Pincovscy[2], Cileno de Magalhães Ribeiro[3], Edna Dias Canedo[4], Fábio Lúcio Lopes de Mendonça[5]

juliano.ferreira@aluno.unb.br; joao.pincovscy@aluno.unb.br; cileno.ribeiro@aluno.unb.br; ednacanedo@unb.br; fabio.mendonca@redes.unb.br

1 2 3 4 5 Universidade de Brasília - UNB, Campus Universitário Darcy Ribeiro, Brasília -DF, CEP: 70910-900, Brasil

Pages: 573-585

Resumo: A Lei Geral de Proteção de Dados Pessoais (LGPD), (aprovada em 2018 no Brasil), obriga a realização da anonimização e a pseudoanonimização para dados pessoais. Diante disso, esse trabalho traz um estudo de caso de privacidade de informações, no qual, se estabelece uma proposta de mitigação dos riscos envolvidos por meio da aplicação de técnicas de anonimização. O objetivo deste trabalho foi enfatizar técnicas de anonimização possíveis, reforçando sua importância na preservação dos dados. Para tanto, foram pesquisados trabalhos nos quais suas peculiaridades na aplicação de técnicas para anonimização atingem sucesso no cumprimento da legislação. Os resultados demonstram a necessidade da seleção correta na técnica de anonimização, visando evitar a possibilidade de identificação dos registros. Como conclusão, se estabeleceu a importância de definir e implementar técnicas de anonimização para proteção de dados, possibilitando salvaguardar a privacidade de pessoas e evitar possíveis ataques de cruzamentos de informação.

Palavras-chave: Anonimização; Pseudoanonimização; Cruzamento de dados; Privacidade.

Mitigation of Privacy Risks after Data Anonymization

Abstract: The General Personal Data Protection Law (LGPD), whichpassed in 2018 in Brazil, requires anonymization and pseudo-anonymization for personal data. Therefore, this work presents a case study of information privacywhere a proposal to mitigate the risks involved is established through the application of anonymization techniques . The objective of this work is to emphasize possible anonymization techniques, reinforcing their importance in data preservation. For that, literature was researched, specifically works with where peculiarities in the application of techniques for anonymization achieve success in complying with the legislation. The results demonstrate the need for the correct selection of the anonymization technique, in order to avoid the possibility of identifying the records. In conclusion,

the importance of defining and implementing anonymization techniques for data protection was established, making it possible to safeguard people's privacy and avoid possible attacks of information crossing.

Keywords: Anonymization; Pseudoanonymization; Data crossing; Privacy.

1. Introdução

A Lei Geral de Proteção de Dados Pessoais (LGPD)(Lei nº 13.709, 2018) aprovada em 2018 e que entrou em vigor em 2020, obriga a anonimização e a pseudoanonimização para dados pessoais. A própria lei define anonimização como "utilização de meios técnicos razoáveis e disponíveis no momento do tratamento, por meio dos quais um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo". A pseudo anonimização é definida como "o tratamento por meio do qual um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo, senão pelo uso de informação adicional mantida separadamente pelo controlador em ambiente controlado e seguro"(Lei nº 13.709, 2018).

Cabe ressaltar que as organizações possuem informações que desejam publicar, mas estas informações podem conter dados pessoais sensíveis (Xue et al., 2011). Em muitos países como no Brasil, Europa e Estados Unidos, os dados médicos são considerados sigilosos desde os primórdios da medicina (Hipócrates, seu patrono, que viveu na grécia por volta de 460 a.C., pregava que a relação entre paciente e médico é secreta (Scarton, [s.d.]). Assim foram desenvolvidos vários métodos e técnicas para realização da anonimização, como sistemas para busca de amostras de bases de dados para execução de pesquisas e trabalhos com a devida anonimização por cifração (Alferes et al., 2011). Utilização de *gateways* para acesso á bases de dados (Bhowmik et al., 2021), utilização da taxonomia em árvore, como método de parcelamento os dados (Rani & Rao, 2021), dentre outros. Nestes estudos os métodos de anonimização para realização de *datamining* (Mineração de Dados com preservação de Privacidade - PPDM), são basicamente de três tipos (Rani & Rao, 2021):

- Método PPDM baseado em perturbação;
- Método PPDM baseado em criptografia;
- Método PPDM baseado em anonimização.

Os métodos baseados em perturbação transformam os dados em uma forma diferente usando técnicas de distorção de dados como adição de ruído, rotação e projeção (Li et al., 2021). Os métodos baseados em criptografia, cifram total ou parcialmente os dados. E os métodos baseados em anonimização transformam os dados substituindo os dados mais específicos por dados menos específicos por generalização para fornecer privacidade (Rani & Rao, 2021). Assim, diante das discrepâncias de conceitos, podemos inferir que a LGPD trata o termo anonimização não como um método e sim como uma forma de ofuscar a informação, impossibilitando a sua associação a um indivíduo.

Toda a problemática dos métodos apresentados é o esforço computacional para modificar as bases de dados e quando necessário recuperá-la de forma íntegra, (Rhahla et al., 2021) além dos possíveis ataques de exfiltração de dados que podem ocorrer mesmo com a adoção dos métodos supracitados (Zheng & Shen, 2021)(Boenisch et al., 2021).

Sendo assim, o objetivo deste trabalho é aplicar técnicas de anonimização para validação em um estudo de caso e avaliar os riscos envolvidos nesse processo de anonimização. Para isso realizamos pesquisas sobre as principais técnicas de anonimização e pseudoanonimização; fizemos a seleção de um subconjunto de dados em uma base de dados hipotética; e aplicamos técnicas de anonimização e pseudoanonimização de dados selecionadas na base de escolhida.

Como contribuição podemos citar que a aplicação adequada de técnicas de anonimização podem resultar na efetiva proteção de dados, reduzindo os riscos associados e mantendo o potencial de utilização da informação. Buscamos ainda apresentar os riscos potenciais à privacidade de dados mesmo após a aplicação de técnicas de anonimização. A inovação reside na discussão do entendimento, inclusive preconizado pela LGPD, de que a simpls aplicação de técnicas de anonimização ou pseudoanonimização de dados seria suficiente para a proteção, porém existem riscos que devem ser considerados.

2. Definições e Trabalhos Correlatos

Nesta seção serão apresentadas as técnicas de anonimização e a aplicação em cenários especficos de utilização das bases de dados.

2.1. Definições

Segundo o Artigo 5 da LGPD/2018, define-se "Anonimização: como a utilização de meios técnicos razoáveis e disponíveis no momento do tratamento, por meio dos quais um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo" (Lei nº 13.709, 2018).

Nos órgãos públicos e privados, segundo a LGPD, é importante que se tenha uma política de governança onde seja indicada a condução do processo de anonimização de dados pessoais e sensíveis (Dias, 2022). Dessa forma, evita-se caso de quebra de privacidade com dados insuficientemente anonimizados, possibilitando que todos os campos de dados confidenciais sejam anonimizados, a fim de remover informações pessoais e reter o que for necessário para análise ou pesquisa posterior.

Pseudonimização é uma técnica de segurança para substituir dados sensíveis por dados fictícios realistas, tendo como definição: "pseudonimização é o tratamento por meio do qual um dado perde a possibilidade de associação, direta ou indireta, a um indivíduo, senão pelo uso de informação adicional mantida separadamente pelo controlador em ambiente controlado e seguro" (Lei nº 13.709, 2018). Destaca-se que ao manter os dados identificados em um ambiente separado, eles só se tornam identificáveis quando os dois elementos são mantidos juntos.

Anonimato ou anonimização é uma técnica que prevê o tratamento dos dados sensíveis para que sejam transformados em dados anonimizados. Assim, por definição "dados anonimizados são os dados relativos a um titular que não possa ser identificado, considerando a utilização de meios técnicos razoáveis e disponíveis na ocasião do seu tratamento" (Cristina & Mahle, 2021). De acordo com essa definição e amparada na LGPD, o titular pode pedir a retirada de dados que possibilitem a sua identificação; os dados vão continuar armazenados, mas não vai mais ser possível identificar a quem

aqueles dados correspondem. Para que essa técnica seja eficiente, ela tem que tornar 'impossível' a identificação do titular daqueles dados (Potiguara Carvalho et al., 2020).

Vale destacar que com o avanço da tecnologia, essa técnica fica cada vez mais passível de reversão, porém a própria lei diz que o mecanismo que for usado no momento dessa anonimização tem que ser considerado razoável e disponível, logo dependente de interpretação. Com a anonimização, os dados são apagados de qualquer informação que possa servir como identificador de uma pessoa, trasnformando dados pessoais em não identificáveis. Essa definição faz com que os dados sejam despojados de qualquer identificação, tornando impossível derivar insights sobre um indivíduo (Cristina & Mahle, 2021).

Mascaramento ou Supressão é uma técnica que utiliza a disponibilização de bases de dados, com informações que não identificam os usuários, porém que pareçam ser reais. As técnicas de mascaramento de dados são: a) substituição: substituição randômica de conteúdo por informações sem relação com o dado real; b) embaralhamento (Shuffing): substituição randômica do dado real por um dado derivado da própria coluna da tabela; c) Blurring: técnica aplicada a números e datas que muda o valor do dado por uma percentagem do seu valor original; e d) Anulação/Truncagem: substitui os dados sensíveis por valor nulos (Azambuja et al., 2019). Assim, a supressão ou mascaramento de dados é uma forma extrema de anonimato. Substitui as informações por algum valor fixo de texto pré-definido (ou em alguns casos, uma tarja preta). Observa-se como vantagem a facilidade de implementar e a eficácia na remoção dos dados. Analogamente, nota-se como desvantagem a perda de qualquer valor estatístico ou analítico dos dados.

Generalização é uma técnica que condiz na substituição dos valores de atributos semi-identificadores por valores menos específicos e com semântica consistente. Assim sendo, os dados são substituídos por valores de categorias mais amplas, por exemplo: o valor 19 do campo "Idade" pode ser substituído por '≤ 20′, o valor '23' por '20 <Idade ≤ 30 ', etc. Dessa forma, percebem-se como vantagens: a possibilidade da análise de dados por categorias além dos dados permanecerem úteis para outras finalidades. Ao passo que se caracteriza como desvantagem o retorno dos dados à sua forma identificável original (Azambuja et al., 2019).

Criptografia homomórfica permite que plataformas de computação em nuvem possam realizar cálculos em dados criptografados homomorficamente sem nunca ter acesso aos dados originais. A implantação da criptografia homomórfica tem uma série de desafios porque esse sistema de criptografia ainda está em desenvolvimento. Suas restrições são baseadas principalmente em como as funções matemáticas são suportadas em dados criptografados (Neto et al., 2020).

Criptografia também é uma técnica que traduz os dados sensíveis em formato criptográfico, sendo necessária uma ou duas chaves secretas para acesso aos dados. Existem dois esquemas de criptografia de dados conhecidos: criptografia simétrica (uma chave) e criptografia assimétrica (duas chaves relacionadas) (Alexandrini & Nardelli, 2021). Em verdade, a escolha de esquemas de criptografia assimétrica, como RSA, DSA ou ECC, que usam duas chaves: pública e privada, aumentam a segurança da informação. (Štarchoň & Pikulík, 2019).

Além dessas técnicas criptográficas, temos a computação segura multipartes, que vem sendo adotada em técnicas de aprendizado de máquina que preservam privacidade; estas permitem a anonimização dos dados usados no treinamento/derivação do modelo, do próprio modelo derivado, e também da consulta (Neto et al., 2020).

Tokenization é uma técnica de abordagem não matemática para proteger dados em repouso que substitui dados sensíveis por substitutos não sensíveis. Essa técnica não altera o tipo ou comprimento dos dados, podendo ser processada por sistemas herdados, tais como bases de dados que podem ser sensíveis ao comprimento e tipo de dados, mantendo a informação sensível oculta. A título de exemplificação a tokenização pode ser usada na segurança dos pagamentos móveis, pois protege as credenciais de pagamento, substituindo-as por um número gerado aleatoriamente que se assemelha ao número de conta primária do cliente, onde apenas o banco e o processador de pagamento podem ler de volta o dado (Štarchoň & Pikulík, 2019).

2.2. Trabalhos Correlatos

As pesquisas nos Estados Unidos sobre o tema são muito úteis, pois abrangem várias situações que só depois da LGPD temos como desafio. No artigo (Gunawan et al., 2021) o autor mostra a importância da anonimização dos dados contidos em uma receita médica, pois existe a correlação direta entre o remédio vendido pelas farmácias e a doença do comprador, apresentando técnicas para ofuscar esta associação. Neste trabalho buscou-se ofuscar os dados de remédios para doenças graves. Por exemplo, na prescrição de Paracetamol não existem dados a serem ofuscados, pois, é um remédio genérico. Por outro lado, a prescrição de uma droga relacionada a problemas do coração devem ser ofuscadas. Pelas amostras estatísticas, o trabalho mostrou que a anonimização destes dados não gera uma perda substancial de dados em pesquisas estatísticas.

Outro fator a se abordado é a utilidade dos dados anonimizados, que em algumas pesquisas podem interferir nos resultados. No artigo (Prata et al., 2020), o modelo de anonimização utilizando o modelo de análise de variância com múltiplos fatores foi aplicado aos dados do Exame Nacional de Desempenho dos Estudantes de graduação no Brasil (ENADE) de 2018. Em uma primeira análise, observou-se que mesmo sendo aplicada a técnica de k-anonimização em meio milhão de registros , ainda teriam um volume muito alto de registros únicos, levando em consideração o código da área do curso, região onde funcionou o curso, idade, gênero, raça/cor e média final do estudante, os níveis de educação da Mãe e do Pai o rendimento do agregado familiar. Assim tiveram que fazer um série de iterações utilizando o Modelo k-anonimato, Modelo ℓ-diversidade e o Modelo t-proximidade para alcançar resultados estatísticos robustos com poucas distorções, trabalhando com dados anonimizados.

Entretanto, de acordo com as pesquisas realizadas, fica evidente que outros métodos de anonimização aplicados a cenários específicos similares podem resultar em sucesso. No artigo (Affonso et al., 2017), utilizou-se como exemplo um conjunto de dados com valores simulados de sujeitos envolvidos em consultas médicas. A base de dados foi estruturada a partir dos elementos da guia de consulta definida no padrão Troca de Informação da Saúde Suplementar – TISS (ANS, 2016), disponibilizada pela Agência Nacional de Saúde – ANS. Primeiramente suprimiram os atributos identificadores

únicos e a generalização dos semi-identificadores que possuem valores do tipo data; em seguida no valor do procedimento foram adicionados valores aleatórios (ruídos) por meio da distribuição normal gaussiana; após vários cálculos matemáticos e estatísticos, chegou-se à conclusão de que ao gerar um conjunto de dados modificados privados para um atributo, é possível preservar a mesma estrutura estatística dos dados originais.

Em outra abordagem, para dados disponibilizados na Internet, por vezes as técnicas de anonimização não são suficientes. No artigo (Potiguara Carvalho et al., 2020), os autores utilizaram dados de fichas financeiras de clientes como amostra inicial. Efetuaram as seguintes estratégias de anonimização: supressão, generalização, agregação, adição de ruídos e substituição. Entretanto, mesmo com a aplicação das técnicas supracitadas, com a utilização de Big Data e acesso a outras fontes de informação, como redes sociais, ainda assim seria possível a identificação de alguns registros.

Um problema observado é a utilidade dos dados após a aplicação de técnicas de anonimização. No artigo (Gunawan, 2020) o autor propôs uma anonimização de dados através do método de substituição. Resultados experimentais mostraram que o método proposto reduz com êxito a probabilidade de sucesso de ataque em um banco de dados anonimizado, ao mesmo tempo em que pode minimizar a perda de informações e preservar as propriedades do banco de dados anonimizado tão semelhante ao original. Embora o método proposto funcione bem na preservação da utilidade dos dados e manutenção das propriedades do banco de dados, deve-se observar que, pela natureza do método de substituição, ele emite distorção de informações.

Após a análise dos trabalhos publicados sobre anonimização observamos que existe um cuidado na técnica de anonimização utilizada, pois pode haver perda de qualidade de acurácia para a pesquisa na utilização dos dados ou até perda de informações importantes, que tornam a base de dados inutilizada.

3. Estudo de Caso

Conforme já contextualizado anteriormente nesse artigo, duas legislações merecem destaque: a European General Data Protection Regulation (GDPR), sendo a legislação baseada na legislação Europeia, e a LGPD, a Lei Geral de Proteção de Dados de agosto de 2018, que é a legislação brasileira a respeito da proteção de dados pessoais. As duas legislações apresentam vários aspectos em comum, em razão do objetivo final de regular a utilização e proteção de dados pessoais ser o mesmo. E merece destaque que ambas legislações apresentam a técnica de Anonimização de dados (Rhahla et al., 2021). Porém a Anonimização de dados não é livre de riscos envolvidos na sua utilização. Existem dois riscos importantes em relação a anonimização de dados, o risco de se perder a capacidade de utilização de dados e o risco de reindentificação desses dados (Canedo et al., 2021).

É importante uma análise adequada das informações armazenadas para reduzir esses riscos envolvidos, tanto no aspecto de utilização do dado como no problema de reindentificação (Piras et al., 2019). E nesse contexto ganha importância a necessidade de uma gestão adequada dessas informações, com a realização de uma governança de dados (Potiguara Carvalho et al., 2020) visando identificar adequadamente quais informações devem ser protegidas, qual técnica de anonimização aplicar e avaliar a manutenção da capacidade de utilização dessa informação.

Considerando um conjunto de informações hipotético, visando aplicação adequada das técnicas de anonimização de dados em busca da proteção dessas informações, a governança de dados preconiza que esses dados sejam analisados e classificados de acordo com suas características e potencial de associação com um indivíduo específico. Dessa maneira, as informações são classificadas em um primeiro tipo de dados que seriam os identificadores de informação pessoal e um segundo tipo de dados que seriam os dados auxiliares, aqueles que não remetem diretamente a identificação de um indivíduo.

Outro aspecto que deve ser considerado nessa classificação das informações é o volume de dados disponível na base de dados trabalhada, pois esse aspecto influencia na capacidade potencial de reindentificação e associação de dados (Carauta Ribeiro & Dias Canedo, 2020), visto que existe um aumento da capacidade inferir determinadas informações a partir de dados disponíveis. A Proteção de Dados Pessoais envolve todos aqueles papéis com interesse na informação, os controladores do dado, processadores (usuários da informação) do dado e o indivíduo a quem o dado se refere (Štarchoň & Pikulík, 2019).

Outra abordagem considerada para a proteção de dados é o uso da técnica de pseudoanonimização de dados (Štarchoň & Pikulík, 2019) sendo uma técnica estruturada na criptografia de dados. Os dados pseudoanonimizados podem ser considerados um subconjunto dos dados pessoais. Com a criptografia dos dados identificadores de informação pessoal, sendo o subconjunto de dados que identifica um indivíduo, e mantendo os dados auxiliares armazenados separadamente, somente é possível atribuir os dados a um indivíduo com informações adicionais ou com a chave criptográfica. Porém, nesse caso, é importante considerar que os dados pseudoanonimizados permitem, de alguma maneira, a reindentificação de um determinado indivíduo (Štarchoň & Pikulík, 2019).

3.1. Hipótese

A avaliação adequada das técnicas de anonimização pode ser efetiva na proteção de dados pessoais sem comprometimento da utilização das informações das bases de dados. Considerando que a seleção de técnicas de anonimização adequadas a cada tipo de dado identificado pode auxiliar no processo de proteção de dados.

3.2. Problema

Como definir quais técnicas de anonimização devem ser utilizadas e de que maneira identificar os dados a serem protegidos, buscando permitir que esses dados ainda possam ser utilizados sem comprometer a privacidade do indivíduo associado a esse dado.

Além disso, deve ser considerado os riscos envolvidos nesse processo de proteção de dados, como o risco de reindentificação das informações e perda de utilidade da informação.

3.3. Objetivo

Buscou-se demonstrar o resultado da anonimização com a possibilidade de utilização parcial das informações sem comprometimento dos dados.

Existe ainda o objetivo de se buscar o alinhamento com as diretrizes definidas nas legislações de proteção de dados.

3.4. Metodologia

Visando avaliar a aplicação de técnicas de anonimização foi selecionado um subconjunto de dados, incluindo informações pessoais para apresentação de um estudo de caso, aplicando a anonimização em dados desanonimizados como Dados de identificação pessoal, sendo aqueles que podem remeter diretamente a identidade de um indivíduo, e por essa razão devem ser protegidos conforme a legislação (Lei nº 13.709, 2018).

A Amostra de dados representada pela Figura 1 foi obtida de fonte aberta, do site da Câmara Legislativa do Brasil (Câmara dos Deputados, 2022). Tratam-se de informações pessoais de Deputados Federais, mas que estão divulgadas abertamente em razão do princípio da transparência. No cenário abordado nesse trabalho, iremos propor a anonimização de determinadas informações pessoais, porém preservando o potencial de utilização do dado (Canedo et al., 2021), ou seja, buscando permitir que a utilização de dados da base apresentada ainda seja possível mesmo com a anonimização de informações.

	id	nomeCivil	cpf	sexo	dataNascimento	ufNascimento	municipioNascimento	escolaridade	uri
1	66179	NORMA AYUB ALVES	28008901187	F	1959-09-07	ES	Vitória	Superior	https://dadosabertos.camara.leg.br/api/v2/deputados/66179
2	66828	FAUSTO RUY PINATO	28022995819	M	1977-06-01	SP	Fernandópolis	Superior	https://dadosabertos.camara.leg.br/api/v2/deputados/66828
3	67138	IRACEMA MARIA PORTELLA NUNES...	37311638372	F	1966-04-23	PI	Teresina	Superior	https://dadosabertos.camara.leg.br/api/v2/deputados/67138
4	68720	FABIO HENRIQUE SANTANA DE ...	41330200578	M	1972-06-19	SE	Simão Dias	Superior	https://dadosabertos.camara.leg.br/api/v2/deputados/68720
5	69871	JOAO CARLOS BACELAR BATISTA	10626409500	M	1957-07-09	BA	Esplanada	Mestrado	https://dadosabertos.camara.leg.br/api/v2/deputados/69871
6	72442	FELIPE AUGUSTO LYRA CARRERAS	86488023420	M	1975-04-16	PE	Recife	Superior ...	https://dadosabertos.camara.leg.br/api/v2/deputados/72442
7	73424	SIMÃO SESSIM	03441067720	M	1935-12-08	RJ	Rio de Janeiro	Superior	https://dadosabertos.camara.leg.br/api/v2/deputados/73424
8	73433	ARLINDO CHIGNALIA JUNIOR	06821146187	M	1949-12-24	SP	Serra Azul	Pós-Graduação	https://dadosabertos.camara.leg.br/api/v2/deputados/73433
9	73441	CELSO UBIRAJARA RUSSOMANNO	01252958803	M	1956-08-20	SP	São Paulo	Superior	https://dadosabertos.camara.leg.br/api/v2/deputados/73441
10	73460	GUSTAVO BONATO FRUET	64446379968	M	1963-04-18	PR	Curitiba	Superior	https://dadosabertos.camara.leg.br/api/v2/deputados/73460
11	73463	OSMAR JOSÉ SERRAGLIO	01773852949	M	1948-05-23	RS	Erechim	Mestrado	https://dadosabertos.camara.leg.br/api/v2/deputados/73463

Figura 1 – Tabela do banco de dados sem anonimização

Observando a Figura 1, verifica-se que são apresentadas de maneira transparente dados que podem ser classificados como dados pessoais, como, nome civil, CPF, sexo, data de nascimento, UF de nascimento, município de nascimento e escolaridade.

Nesse contexto cabem duas categorias de classificação para essas informações: a primeira como dados de identificação pessoal e dados auxiliares (Canedo et al., 2021), sendo que a primeira categoria remete diretamente à identificação do indivíduo e os dados auxiliares não identificam de maneira isolada o indivíduo, porém podem auxiliar nessa identificação.

Na primeira categoria de dados de identificação pessoal podem ser incluídos o Nome Civil, CPF e data de nascimento. Na segunda categoria, como dados auxiliares podem ser indicados o sexo, UF de nascimento, município de nascimento e escolaridade.

Dessa maneira uma possibilidade de anonimização dessas informações buscando proteger a privacidade dos dados pessoais, com a aplicação de técnicas de anonimização nos dados classificados como dados de identificação pessoal.

Em nosso exemplo foi utilizada a ferramenta Excel para o mascaramento, substituição e supressão de informações. Essa ferramenta foi utilizada a título de exemplo, apenas para simular a aplicação dessas técnicas, não foram utilizados algoritmos de aplicação automática dessas técnicas, mas sim a ferramenta com o objetivo de simular e demonstrar o resultado dos dados anonimizados.

ID	Nome Civil	CPF	SEXO	Data de Nascimento	Uf Nascimento	Municipio	Escolaridade
66179	NORMA AYUB ALVES	28008901187	F	07/09/1959	ES	Vitória	Superior
66828	FAUSTO RUY PINATO	28022995819	M	01/06/1977	SP	Fernandópolis	Superior
67138	IRACEMA MARIA PORTELLA NUNES NOGUEIRA LIMA	37311638372	F	23/04/1966	PI	Teresina	Superior
68720	FABIO HENRIQUE SANTANA DE CARVALHO	41330200578	M	19/06/1972	SE	Simão Dias	Superior

Figura 2 – Base de dados antes da anonimização.

3.5. Resultados

Na Figura 2 é apresentado um subconjunto de dados listados na Figura 1 a título de exemplo de como os dados são apresentados antes do processo de anonimização. Considerando os dados pessoais apresentados na figura 2 é possível selecionar técnicas de anonimização a serem utilizadas.

Para o dado pessoal denominado Nome Civil a proposta para estudo de caso é de aplicação da técnica de supressão (Canedo et al., 2021) que implica na exclusão dessa informação da base de dados. Essa técnica é uma opção quando a informação tem um grau de associação com o indivíduo muito grande e quando a informação não é necessária para a obtenção de outras informações como consolidações estatísticas.

Para o dado pessoal CPF, a proposta nesse estudo de caso é a aplicação da técnica de mascaramento, em que parte significativa das informações são mascaradas por caractere pré-definido. Essa técnica permite uma identificação parcial do registro, porém não permite a associação direta com o indivíduo.

Por último, em relação à data de nascimento, a técnica aplicada é a de substituição, que trata do embaralhamento dessas informações na base de dados, fazendo uma troca das informações na base de dados.

Assim é possível ter a utilidade das informações em relação à data de nascimento dos indivíduos cadastrados, porém sem a associação direta a identificação ou reidentificação desse indivíduo.

1	ID	CPF	SEXO	Data de Nascimento	Uf Nascimento	Municipio	Escolaridade
2	66179	XXXXXXXX187	F	23/04/1966	ES	Vitória	Superior
3	66828	XXXXXXXX819	M	19/06/1972	SP	Fernandópolis	Superior
4	67138	XXXXXXXX372	F	07/09/1959	PI	Teresina	Superior
5	68720	XXXXXXXX578	M	01/06/1977	SE	Simão Dias	Superior

Figura 3 – Base de dados anonimizada.

Observando a Figura 3, é possível verificar o resultado das aplicações das técnicas de anonimização propostas.

3.6. Discussão

O resultado permite ainda a utilidade de informações como Sexo, Data de nascimento, UF e municípios de nascimento e, ainda, escolaridade. Porém, esse processo evita a imediata identificação dos indivíduos cadastrados nessa base de dados.

Cabe, no entanto, uma observação adicional dos riscos envolvidos no processo de anonimização apresentado nesse estudo de caso, como dados auxiliares não foram anonimizados. É importante o registro da ameaça de reidentificação dos dados (Štarchoň & Pikulík, 2019) quando da associação dessas informações com outros dados obtidos em fontes abertas, como seria o caso de consulta em fonte aberta viabilizar a associação da data de nascimento, uf e do município de nascimento com a identidade do indivíduo.

4. Conclusões

Analisando a LGPD, observamos obviamente que a legislação tem grande preocupação com os dados dos indivíduos obrigando a anonimização. Entretanto, quando nos aprofundamos nas questões técnicas, nos deparamos com situações que, a princípio, atenderiam plenamente a legislação, mas diante de uma análise mais aprofundada descobrimos que a realidade não é exatamente esta.

Primeiramente, já nos deparamos com diferença de definições no que se refere ao termo anonimização, pois no meio acadêmico técnico, o termo é considerado um dos métodos utilizados para preservação de privacidade de dados. Assim, seria mais apropriado utilizar o termo "privacidade de dados" e não "anonimização".

Outra descoberta relevante em nossa pesquisa foi que existe uma vasta pesquisa em privacidade de dados, com inúmeras técnicas para a "anonimização", e quase todas realmente efetivas. Entretanto, cada técnica utilizada se adéqua a um cenário específico para ser realmente efetiva e eficiente no que se refere à utilização dos dados.

Em nosso estudo de caso ficou evidente que a utilização da bases de dados anonimizada em pesquisas controladas é extremamente útil e eficiente, atendendo com êxito a LGPD. Como utilizamos técnicas simples de ofuscação de informações, esta base anonimizada deve ser utilizada em ambientes de redes controladas. Caso fôssemos expor as informações na Internet, esta estaria sujeita à reindentificação do indivíduo efetuando o cruzamento de informações de outras fontes de dados.

Essa ameaça pode ser minimizada com aplicação de técnicas de anonimização adicionais nessas informações (Štarchoň & Pikulík, 2019), porém é importante a avaliação de cenário no caso concreto para evitar a redução significativa de utilidade do dado com a anonimização mais ampla das informações (Canedo et al., 2021).

Os ataques à "anonimização" são um fator determinante neste tema. Apesar de as pesquisas nesta área estarem avançadas, os ataques também estão cada vez mais sofisticados. Aliadas à sofisticação temos ainda as capacidades computacionais e a vastidão de fontes de dados abertas na Internet, que possibilitam cruzamentos infinitos de registros, tornando os ataques mais eficientes.

Através do estudo de caso apresentado, outros atores podem utilizar técnicas similares às apresentadas nesse estudo para avançar na abordagem de proteção de dados.

Neste trabalho não fizemos uso de técnicas de Aprendizado de Máquinas e da Inteligência Artificial. Entretanto, fica para reflexão o cenário atual onde técnicas de IA estão cada vez mais presentes na governança dos dados necessitaremos de um cuidado cada vez maior para anonimizar as informações.

Agradecimentos

Os autores agradecem o suporte da ABIN TED 08/2019, além do CNPq, CAPES e Câmara dos Deputados do Brasil pela disponibilização das bases de dados e pesquisa.

Referências

Affonso, E. P., De Oliveira, S. C., & Sant'Ana, R. C. G. (2017). Análise do equilíbrio entre privacidade e utilidade no acesso a dados. *Informacao e Sociedade*, *27*(1), 81–92. https://doi.org/10.22478/ufpb.1809-4783.2017v27n1.29422

Alexandrini, F., & Nardelli, C. (2021, dezembro). Primeira Fase da Segurança da Informação e LGPD Aplicado no Desenvolvimento de Software Governo Eletrônico. *Revista de Extensão e Iniciação Científica da UNISOCIESC*, 20. https://doi.org/10.13140/RG.2.2.30059.46883

Azambuja, A. J. G. de, Granville, L. Z., & Sarmento, A. G. M. (2019). A privacidade, a segurança da informação e a proteção de dados no Big Data. *Parcerias Estratégicas*, 9–32. https://tinyurl.com/26vwz3cx

Bhowmik, C., Momin, M. A. I., Shanta, F. J., & Haque, M. M. (2021). Database Security as a Gateway to Privacy Preserving Data Mining. *International Conference on Robotics, Electrical and Signal Processing Techniques*, 125–130. https://doi.org/10.1109/ICREST51555.2021.9331190

Boenisch, F., Munz, R., Tiepelt, M., Hanisch, S., Kuhn, C., & Francis, P. (2021). Side-Channel Attacks on Query-Based Data Anonymization. In *Proceedings of the ACM Conference on Computer and Communications Security* (Vol. 1, Número 1). Association for Computing Machinery. https://doi.org/10.1145/3460120.3484751

Câmara dos Deputados. (2022). *Dados Abertos da Câmara dos Deputados*. https://dadosabertos.camara.leg.br/

Canedo, E., Cerqueira, A., Gravina, R., Ribeiro, V., Camões, R., Reis, V., Mendonça, F., & Sousa Jr., R. (2021). *Proposal of an Implementation Process for the Brazilian General Data Protection Law (LGPD)*. 19–30. https://doi.org/10.5220/0010398200190030

Carauta Ribeiro, R., & Dias Canedo, E. (2020). Using MCDA for Selecting Criteria of LGPD Compliant Personal Data Security. *The 21st Annual International Conference on Digital Government Research*, 175–184. https://doi.org/10.1145/3396956.3398252

Cristina, A., & Mahle, O. (2021). *A Autodeterminação Informativa como Fundamento da Lei Geral de Proteção de Dados Brasileira: Uma Análise a Partir da LGPD* [Universidade Federal de São Carlos]. https://repositorio.ufscar.br/bitstream/handle/ufscar/15630/Ana Cristina Oliveira Mahle essa.pdf?sequence=1&isAllowed=y

Dias, S. P. (2022). *Autonomia Heteronoma e Discriminação Algorítmica: Análise do piso normativo para admissão do tratamento discriminatório de dados pessoais do titular* [Universidade Federal de Ouro Preto]. http://shorturl.at/qvKTW

Gunawan, D. (2020). A Data Anonymization Method to Mitigate Identity Attack in Transactional Database Publishing. *2020 8th International Conference on Information and Communication Technology, ICoICT 2020*. https://doi.org/10.1109/ICoICT49345.2020.9166262

Gunawan, D., Nugroho, Y. S., Maryam, & Al Irsyadi, F. Y. (2021). Anonymizing Prescription Data against Individual Privacy Breach in Healthcare Database. *2021 9th International Conference on Information and Communication Technology, ICoICT 2021*, 138–143. https://doi.org/10.1109/ICoICT52021.2021.9527430

José Alferes, D., Miguel Goulão, D., José Alberto Cardoso Cunha Arguente, D., Vitor Manuel Beires Pinto Nogueira Vogal, D., & José Júlio Alves Alferes, D. (2011). *Building Anonymised Database Samples Dissertação para obtenção do Grau de Mestre em Engenharia Informática*. Universidade Nova de Lisboa.

Lei nº 13.709, de 14 de A. de 2018. (2018). *Lei Geral de Proteção de Dados Pessoais (LGPD)*. Legislação Brasileira. http://shorturl.at/tzCW9

Li, J., Han, L., Chen, R., Zhang, H., Han, B., Wang, L., & Cao, X. (2021). Identity-Preserving Face Anonymization via Adaptively Facial Attributes Obfuscation. *MM 2021 - Proceedings of the 29th ACM International Conference on Multimedia*, 3891–3899. https://doi.org/10.1145/3474085.3475367

Neto, H. N. C., Mattos, D. M. F., & Fernandes, N. C. (2020). Privacidade do Usuário em Aprendizado Colaborativo: Federated Learning, da Teoria à Prática. In E. Viefas (Org.), *Minicursos do SBSeg* (p. 101-). Simpósio Brasileiro de Segurança da Informação e de Sistemas Computacionais. https://www.lgpdbrasil.com.br/.

Piras, L., Al-Obeidallah, M. G., Praitano, A., Tsohou, A., Mouratidis, H., Gallego-Nicasio Crespo, B., Bernard, J. B., Fiorani, M., Magkos, E., Sanz, A. C., Pavlidis, M., D'Addario, R., & Zorzino, G. G. (2019). DEFeND Architecture: A Privacy by Design Platform for GDPR Compliance. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11711 LNCS*, 78–93. https://doi.org/10.1007/978-3-030-27813-7_6

Potiguara Carvalho, A., Potiguara Carvalho, F., Dias Canedo, E., & Potiguara Carvalho, P. H. (2020). Big data, anonymisation and governance to personal data protection. *ACM International Conference Proceeding Series*, 185–195. https://doi.org/10.1145/3396956.3398253

Prata, P., Ferrão, M. E., Santos, W., & Sousa, G. (2020). Privacy preserving versus utility preserving in data anonymization: A study in higher education. *RISTI - Revista Iberica de Sistemas e Tecnologias de Informacao*, *2020*(E40), 112–127. https://doi.org/10.17013/RISTI.40.112-127

Rani, V. U., & Rao, M. S. (2021). PrivGuard: Sensitivity Guided Anonymization based PPDM with Automatic Selection of Sensitive Attributes. *2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021*, 1832–1837. https://doi.org/10.1109/ICACCS51430.2021.9441991

Rhahla, M., Allegue, S., & Abdellatif, T. (2021). Guidelines for GDPR compliance in Big Data systems. *Journal of Information Security and Applications*, *61*(June), 102896. https://doi.org/10.1016/j.jisa.2021.102896

Scarton, R. R. ([s.d.]). *Violação do Segredo Profissional Dos Médicos: Aspectos Jurídicos e (Bio)Éticos - Âmbito Jurídico - Educação jurídica gratuita e de qualidade*. Recuperado 31 de maio de 2022, de http://shorturl.at/eCLN1

Štarchoň, P., & Pikulík, T. (2019). GDPR principles in Data protection encourage pseudonymization through most popular and full-personalized devices - mobile phones. *Procedia Computer Science*, *151*, 303–312. https://doi.org/10.1016/J.PROCS.2019.04.043

Xue, M., Karras, P., Raïssi, C., & Pung, H. K. (2011). Utility-driven anonymization in data publishing. *International Conference on Information and Knowledge Management, Proceedings*, 2277–2280. https://doi.org/10.1145/2063576.2063945

Zheng, J., & Shen, X. (2021). Pattern mining and detection of malicious sql queries on anonymization mechanism. *IEEE Access*, *9*, 15015–15027. https://doi.org/10.1109/ACCESS.2021.3052956